

\mathcal{M} -DECOMPOSABILITY, ELLIPTICAL UNIMODAL DENSITIES, AND APPLICATIONS TO CLUSTERING AND KERNEL DENSITY ESTIMATION

NICHOLAS CHIA AND JUNJI NAKANO

ABSTRACT. Chia and Nakano (2009) introduced the concept of \mathcal{M} -decomposability of probability densities in one-dimension. In this paper, we generalize \mathcal{M} -decomposability to any dimension. We prove that all elliptical unimodal densities are \mathcal{M} -undecomposable. We also derive an inequality to show that it is better to represent an \mathcal{M} -decomposable density via a mixture of unimodal densities. Finally, we demonstrate the application of \mathcal{M} -decomposability to clustering and kernel density estimation, using real and simulated data. Our results show that \mathcal{M} -decomposability can be used as a non-parametric criterion to locate modes in probability densities.

1. INTRODUCTION

In a recent paper, Chia and Nakano (2009) conceptualized \mathcal{M} -decomposability and developed the theory in one-dimension. The main results are summarized in the following paragraph.

\mathcal{M} -decomposability is defined as follows. Let f be a probability density defined in one-dimension. There exist countless ways to express f as a weighted mixture of two probability densities, in the form of

$$f(x) = \alpha g(x) + (1 - \alpha) h(x) \quad \text{where } 0 < \alpha < 1.$$

If it is possible to find any combination of $\{\alpha, g, h\}$, which satisfies

$$\sigma_f > \sigma_g + \sigma_h \quad \text{where } \sigma_f \text{ denotes the standard deviation of } f,$$

then the original density f is said to be \mathcal{M} -decomposable. Otherwise, f is \mathcal{M} -undecomposable. Intuitively, multimodal densities with peaks separated far apart are likely to be \mathcal{M} -decomposable. Conversely, unimodal densities are probably \mathcal{M} -undecomposable. The authors proved that all one-dimensional *symmetric unimodal densities* with finite second moments are \mathcal{M} -undecomposable. In other words, if f is symmetric unimodal and has finite second moments, then for any weighted mixture density components $\{g, h\}$ of f , one must have

$$(1.1) \quad \sigma_f \leq \sigma_g + \sigma_h.$$

Eq (1.1) applies to a wide range of densities that include Gaussian, Laplace, logistic and many others. The authors also showed the possibility of using \mathcal{M} -decomposability to perform cluster analysis and mode finding in one-dimension.

2000 *Mathematics Subject Classification.* Primary 62H30, 62G07; Secondary 15A45.

Key words and phrases. covariance matrices, inequalities, cluster analysis, elliptical unimodal densities, Kullback-Leibler divergence, density estimation, non-parametric criterion.

Incidentally, the “ \mathcal{M} ” in \mathcal{M} -decomposability may either mean “multimodal” or “mixture”.

In this paper, we further contribute to \mathcal{M} -decomposability, both in the theoretical and applicational aspects. On the theoretical front, we generalize the concept of \mathcal{M} -decomposability to any d -dimensional space. First of all, we derive a theorem (Theorem 2.3) that is the d -dimensional equivalent of Eq (1.1). We prove that all *elliptical unimodal densities* with finite second moments are \mathcal{M} -undecomposable. These densities include multivariate Gaussian, Laplace, logistic and many others. Following that, we derive another theorem, (Theorem 2.4), which determines if a given density is better approximated via a mixture of Gaussian densities, instead of one single Gaussian density.

One example of application of \mathcal{M} -undecomposability is cluster analysis. For decades, cluster analysis has been a popular research subject, both from the theoretical and algorithmic aspects. Cluster analysis is likely to remain a widely researched topic, given the many different approaches that caters to varying applications. The survey paper by Berkhin (2002) provides an up-to-date status of available clustering techniques and methodologies. There are two main classes of cluster analysis methodologies: *parametric* and *non-parametric*. For parametric cluster analysis, one needs prior knowledge or assumptions on the analytical structure of the underlying clusters. The whole dataset is modeled as a mixture of k parametrized densities, and the problem reduces to parameter estimation. In McLachlan and Peel (2000), parametric cluster analysis via the *Expectation-Maximization* (EM) algorithm is described in detail. Other parametric methods include the Bayesian particle filter approach detailed in Fearnhead (2004), and the reversible jump *Markov chain Monte Carlo* (MCMC) approach by Richardson and Green (1997). For parametric cluster analysis, the most popular approach is to model the clusters as Gaussian densities.

As for non-parametric cluster analysis, a popular tool is the k -means algorithm. The k -means algorithm is optimal for locating similar-sized spherical clusters within a dataset, provided the number of clusters are known beforehand. With elliptical clusters, or clusters of varying sizes, the k -means approach yields results that are meaningless. The k -means algorithm assigns samples to clusters based on *distance* (Euclidean or its variations) to the centres of the clusters. Other *distance-based* non-parametric clustering algorithms include the nearest-neighbour clustering. Distance-based clustering algorithms generally share the same drawbacks such as sensitivity to scaling, elliptical clusters and clusters of varying sizes. If the number of clusters are not known beforehand, neither the k -means algorithm nor the nearest-neighbour algorithm estimate the number of clusters automatically. For the k -means algorithm, the unknown number of clusters has to be re-evaluated via *Akaike’s information criterion* (AIC), proposed by Akaike (1974), or other suitable model selection criterion.

Our approach to cluster analysis via \mathcal{M} -decomposability is non-parametric and are based on *volume* instead of distance. Being non-parametric, prior knowledge on the analytical structure of the underlying clusters is unnecessary. The only assumption required is that the clusters are approximately elliptical and unimodal. As a result, the limitation of clustering via \mathcal{M} -undecomposability is that it will probably not perform ideally for irregularly shaped clusters that deviate from elliptical unimodal densities. However, if the clusters are approximate elliptical and

unimodal, then our clustering methodology works well, and allows for the unknown number of clusters to be recovered automatically. Furthermore, as clustering via \mathcal{M} -decomposability is based on volume instead of distance, cluster allocation is invariant to scaling.

For existing alternative methodologies to clustering, there has been recent development on Rousseeuw’s minimum volume ellipsoids (MVE) in Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990). The MVE approach is originally developed as a robust method to estimate mean vectors and covariance matrices of multivariate data in the presence of outliers. MVE is computationally intensive and the optimal solution is often difficult to achieve, prompting many research papers on the algorithmic aspects of the problem. Some authors, for example, Shioda and Tünel (2005), outlined a heuristic for clustering via MVE by minimizing the sum of volume of clusters. Our methodology of clustering via \mathcal{M} -decomposability has some similarities with clustering via the MVE approach, in that both measure “volume” in a certain sense. Central to the \mathcal{M} -decomposability concept is the “pseudo-volume”, which we define as the square-root of the determinant of the covariance matrix. Compared to MVE, the pseudo-volume is computationally cheap and straightforward. On top of that, we also provide theoretical justifications in Theorem 2.4 for minimizing the sum of pseudo-volumes of clusters.

Another possible area of application of \mathcal{M} -undecomposability is density estimation. In density estimation, data generated from some unknown densities are given, and the task is to estimate and recover the unknown density. One popular non-parametric approach to density estimation is kernel density estimation, treated in Silverman (1986), Scott (1992), Härdle *et al* (2004), as well as Wand and Jones (1995). The difficulty in kernel density estimation is the derivation of the optimal kernel bandwidth: If the kernel bandwidth is underestimated, the kernel density becomes unduly spiky; if the kernel bandwidth is overestimated, the kernel density becomes oversmoothed. For multimodal densities, it is not possible to find a single kernel bandwidth that provides a satisfactory density estimation everywhere. Using \mathcal{M} -decomposability, we demonstrate that there is a simple and logical way to circumvent the above problem by representing the underlying density as a mixture of unimodal densities where necessary.

This paper develops both the theoretical and applicational aspects of \mathcal{M} -decomposability, and therefore should be of interest to theoretical statisticians and practitioners alike. Section 2 is devoted to the theoretical development of \mathcal{M} -decomposability in d -dimensional space. For readers who are only interested in applications, it is possible to note only the results of Theorems 2.3 and 2.4, skipping the rest of Section 2 without disrupting the flow of the paper.

2. \mathcal{M} -DECOMPOSABILITY IN d -DIMENSIONAL SPACE

2.1. Extensions from One-Dimension. In Chia and Nakano (2009), \mathcal{M} -decomposability involves only the standard deviations of probability densities. This is because in one-dimension, the standard deviation is a natural measure of scatter of a given density. The standard deviation of any density in one-dimension has the same order as the distance or “length” computed from the mean. When considering higher dimensions, a possible corresponding measure of scatter of a given density is the square-root of the determinant of the covariance matrix of the density. The square-root of the determinant of the covariance matrix in d -dimensional space

has the same order as d -dimensional “hypervolume”. Henceforth, we shall call the above measure the *pseudo-volume* of a density. We denote the covariance matrix of a density f by Σ_f , and therefore the pseudo-volume of f is given by $|\Sigma_f|^{\frac{1}{2}}$. In one-dimension, pseudo-volume reduces to the standard deviation.

In Chia and Nakano (2009), the authors limited the number of mixture components to two in their development of \mathcal{M} -decomposability. In this paper, we show that it is possible to relax the above limitation, and generalize the number of mixture components to m where $m \geq 2$. Let f be a probability density function defined on \mathcal{R}^d , the d -dimensional real space. One can always express f as a weighted mixture of m densities as follows:

$$(2.1) \quad f(\mathbf{x}) = \alpha_1 g_1(\mathbf{x}) + \cdots + \alpha_m g_m(\mathbf{x}),$$

where $0 < \alpha_i < 1$ and $\sum \alpha_i = 1$. Henceforth, we call any set of densities $\{g_1, \dots, g_m\}$ which satisfies Eq (2.1) a set of *mixture components* of f .

We extend the definition of \mathcal{M} -decomposability to d -dimensional space as follows.

Definition 2.1 (\mathcal{M} -Decomposability). *For a given probability density function f , if there exists a set of mixture components $\{g_1, \dots, g_m\}$ such that*

$$|\Sigma_f|^{\frac{1}{2}} > |\Sigma_{g_1}|^{\frac{1}{2}} + \dots + |\Sigma_{g_m}|^{\frac{1}{2}},$$

then f is defined to be \mathcal{M} -decomposable. Otherwise, f is \mathcal{M} -undecomposable. If for any set of mixture components $\{g_1, \dots, g_m\}$,

$$|\Sigma_f|^{\frac{1}{2}} < |\Sigma_{g_1}|^{\frac{1}{2}} + \dots + |\Sigma_{g_m}|^{\frac{1}{2}},$$

then f is strictly \mathcal{M} -undecomposable.

Our new definition of \mathcal{M} -decomposability reduces to that presented in Chia and Nakano (2009) when $m = 2$ and $d = 1$. For $d \geq 2$, the definition of \mathcal{M} -decomposability can be described compactly using pseudo-volumes.

2.2. Elliptical Uniform Densities. The uniform density is trivially defined in one-dimension, but in higher dimensions, it may assume many different possible shapes. For example, one may think of the uniform hypercube or the uniform hypersphere. However, the subject of interest in our paper is the *elliptical uniform density*, which forms the fundamental building block of elliptical unimodal densities.

Ellipticity, uniformity and unimodality are three different qualities. The definitions of the first two are given immediately below, and the third will be given in Section 2.3.

Definition 2.2 (Elliptical and Spherical Densities). *We say that f is elliptical if there exist a vector $\mu \in \mathcal{R}^d$, a positive semidefinite symmetric matrix $\Sigma \in \mathcal{R}^{d \times d}$ and a positive function p on $\mathcal{R}^+ \cup \{0\}$ such that*

$$f(\mathbf{x}) = p\{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\}.$$

Furthermore, if $\Sigma = k \mathbf{I}_d$, where $k > 0$ and \mathbf{I}_d denotes the d -dimensional identity matrix, then f becomes

$$f(\mathbf{x}) = p_1\{(\mathbf{x} - \mu)^T (\mathbf{x} - \mu)\} = p_2(|\mathbf{x} - \mu|),$$

and we say that f is spherical.

Remark The mean and covariance matrix of the above-defined elliptical density f are as follows:

$$\mu_f = \mu, \quad \Sigma_f = c \Sigma \quad \text{where } c > 0.$$

Definition 2.3 (Uniform Densities). *We say that f is elliptical uniform if there exist a vector $\mu \in \mathcal{R}^d$, a positive semidefinite symmetric matrix $\Sigma \in \mathcal{R}^{d \times d}$, and a positive real number r such that*

$$f(\mathbf{x}) \propto \mathbb{I}_{(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) < r^2},$$

where \mathbb{I} denotes the indicator function. Furthermore, if $\Sigma = k \mathbf{I}_d$, where $k > 0$ and \mathbf{I}_d denotes the d -dimensional identity matrix, then f becomes

$$f(\mathbf{x}) \propto \mathbb{I}_{(\mathbf{x}-\mu)^T (\mathbf{x}-\mu) < r'^2} = \mathbb{I}_{|\mathbf{x}-\mu| < r'},$$

and we say that f is spherical uniform.

Theorem 2.1 (Inequality on Elliptical Uniform Densities). *All elliptical uniform densities defined on \mathcal{R}^d are \mathcal{M} -undecomposable in $d = 1$ and strictly \mathcal{M} -undecomposable for $d \geq 2$.*

The proof of Theorem 2.1 proceeds the following lemma.

Lemma 2.1 (Density with Minimum Pseudo-volume). *Let f be a probability density function defined on $\mathbf{x} \in \mathcal{R}^d$ such that $f(\mathbf{x}) \leq M_f$ for all \mathbf{x} . Then*

$$|\Sigma_f|^{\frac{1}{2}} \geq \frac{\Gamma(\frac{d}{2} + 1)}{M_f \{\pi(d+2)\}^{\frac{d}{2}}}.$$

Identity holds if and only if f is elliptical uniform with $\max(f) = M_f$.

Remark When $d = 1$, we recover $\sigma_f \geq 1/(M_f \sqrt{12})$, the result obtained in Chia and Nakano (2009).

The proof of the Lemma 2.1 has been relegated to Section 5.2 of the appendix to enhance the flow of the paper. We use the results of Lemma 2.1 to prove Theorem 2.1.

Proof of Theorem 2.1. Let u be an elliptical uniform density on $\mathbf{x} \in \mathcal{R}^d$ ($d \geq 1$). We need to prove that for any set of mixture components $\{v_1, \dots, v_m\}$ of u ,

$$|\Sigma_{v_1}|^{\frac{1}{2}} + \dots + |\Sigma_{v_m}|^{\frac{1}{2}} > |\Sigma_u|^{\frac{1}{2}}.$$

Without loss of generality, set $\max(u) = M$ and therefore

$$|\Sigma_u|^{\frac{1}{2}} = \frac{\Gamma(\frac{d}{2} + 1)}{M \{\pi(d+2)\}^{\frac{d}{2}}}.$$

Rewriting the elliptical uniform density u as mixture components, we have

$$u(\mathbf{x}) = \alpha_1 v_1(\mathbf{x}) + \dots + \alpha_m v_m(\mathbf{x})$$

for some $\{\alpha_1, \dots, \alpha_m\}$ satisfying $0 \leq \alpha_j \leq 1$ and $\sum \alpha_j = 1$. As a result, we have

$$v_j(\mathbf{x}) \leq \frac{u(\mathbf{x})}{\alpha_j} \leq \frac{M}{\alpha_j}$$

for all $1 \leq j \leq m$. Using Lemma 2.1, we have

$$(2.2) \quad |\Sigma_{v_j}|^{\frac{1}{2}} \geq \frac{\alpha_j \Gamma(\frac{d}{2} + 1)}{M \{\pi(d+2)\}^{\frac{d}{2}}} = \alpha_j |\Sigma_u|^{\frac{1}{2}}$$

for all j , with equalities holding if and only if the density in question is elliptical uniform. Now, for $d > 1$, we can have *at most* $(m - 1)$ but *never all* of v 's to be elliptical uniform satisfying Eq (2.2). Therefore,

$$|\Sigma_{v_1}|^{\frac{1}{2}} + \dots + |\Sigma_{v_m}|^{\frac{1}{2}} \geq |\Sigma_u|^{\frac{1}{2}}.$$

Identity may only hold when $d = 1$, refer to Chia and Nakano (2009). \square

2.3. Elliptical Unimodal Densities. In one-dimension, symmetry is trivial to visualize and express mathematically. In higher dimensions, symmetry may be depicted via ellipticity. As such, *elliptical unimodal densities* play a key role in this paper. We provide a definition for elliptical unimodal densities below. Elliptical densities in general have been treated in detail by many researchers, see Fang *et al* (1990) and references within. Unimodal densities have also been the subject of active research. For example, refer to Anderson (1955), Dharmadhikari and Joag-Dev (1987) as well as Ibragimov (1956).

Definition 2.4 (Elliptical Unimodal Densities). *We say that f is elliptical unimodal if there exist a vector $\mu \in \mathcal{R}^d$, a positive semidefinite symmetric matrix $\Sigma \in \mathcal{R}^{d \times d}$ and a non-increasing positive function p on $\mathcal{R}^+ \cup \{0\}$ such that*

$$f(\mathbf{x}) = p\{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\}.$$

Comparing with Definition 2.2, the only additional information in Definition 2.4 is that the positive function p has to be non-increasing as well. According to Definition 2.4, elliptical unimodal densities are those whose cross-sections are elliptical, and with mean (μ) and covariance matrices proportional to (Σ). Definition 2.4 encompasses a large class of general densities including d -dimensional elliptical uniform, Gaussian, logistic, Laplace, Von Mises, beta(k, k) where $k > 1$, student- t , and many other densities.

Henceforth, we propose the following alternative representation of elliptical unimodal densities.

Theorem 2.2 (Representation of Elliptical Unimodal Densities). *Let f be an elliptical unimodal density with mean μ and covariance matrix Σ . Then, for all $\epsilon > 0$, it is possible to construct a density*

$$g_n(\mathbf{x}) = \sum_{j=1}^n b_j u_j(\mathbf{x})$$

such that

$$\int |g_n(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} < \epsilon.$$

Here, each u_j is an elliptical uniform density such that

$$(2.3) \quad u_j(\mathbf{x}) \propto \mathbb{I}_{(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) < r_j^2}$$

and r 's are strictly positive. Furthermore, each proportionality constant b_j satisfies

$$b_j = \frac{r_j^d}{\sum_{i=1}^n r_i^d}.$$

From the above representation, each elliptical uniform component is weighted proportionally to the hypervolume of its cross-section. The original elliptical unimodal density is “sliced latitudinally” into elliptical uniforms with a prefixed constant “thickness”. The proof of Theorem 2.2 has been relegated to Section 5.3 of the appendix.

2.4. A Theorem on Elliptical Unimodal Densities.

Theorem 2.3 (Inequality on Elliptical Unimodal Densities). *Let f be an elliptical unimodal density with finite second moments. Then, for any set of mixture components $\{g_1, \dots, g_m\}$,*

$$|\Sigma_f|^{\frac{1}{2}} \leq |\Sigma_{g_1}|^{\frac{1}{2}} + \dots + |\Sigma_{g_m}|^{\frac{1}{2}}.$$

Identity is possible only when f is uniform in one-dimension.

Proof. Our task is to prove that for all mixture components $\{g_1, \dots, g_m\}$ satisfying

$$(2.4) \quad f(\mathbf{x}) = \sum_{i=1}^m a_i g_i(\mathbf{x}),$$

where $0 < a_i < 1$ and $\Sigma a_i = 1$, we must have

$$(Claim\ 1) \quad |\Sigma_f|^{\frac{1}{2}} \leq |\Sigma_{g_1}|^{\frac{1}{2}} + \dots + |\Sigma_{g_m}|^{\frac{1}{2}}.$$

Using Theorem 2.2, we can approximate f to an arbitrary level of accuracy by rewriting f as a finite mixture of elliptical uniform densities, each having “uniform thickness” as

$$(2.5) \quad f(\mathbf{x}) = \sum_{j=1}^n b_j u_j(\mathbf{x}).$$

The “thickness” of each elliptical uniform component is equal to $\max\{b_j u_j(\mathbf{x})\}$. Here, u_j ’s, as described in Eq (2.3), are elliptical uniform densities sharing the same means and whose covariances are multiples of each other. Each constant of proportionality, denoted by b_j , is proportional to the hypervolume of the corresponding elliptical uniform density u_j .

To provide a link between Eqs (2.4) and (2.5), we further rewrite f as

$$f(\mathbf{x}) = \sum_{i=1}^m \sum_{j=1}^n c_{i,j} v_{i,j}(\mathbf{x}) = \sum_{i=1}^m a_i g_i(\mathbf{x}) = \sum_{j=1}^n b_j u_j(\mathbf{x}).$$

For each pair of (i, j) above, $c_{i,j} v_{i,j}(\mathbf{x})$ is the “intersection” of the segments $a_i g_i(\mathbf{x})$ and $b_j u_j(\mathbf{x})$ with respect to f on the curve. For all values of $\{i, j\}$, g_i and u_j can be expressed in terms of $v_{i,j}$ as

$$(2.6) \quad a_i g_i(\mathbf{x}) = \sum_{j=1}^n c_{i,j} v_{i,j}(\mathbf{x}), \quad b_j u_j(\mathbf{x}) = \sum_{i=1}^m c_{i,j} v_{i,j}(\mathbf{x}).$$

Here, depending on the mixture components $\{g_1, \dots, g_m\}$, it is possible for some of $c_{i,j}$ ’s to be 0, as long as for all values of $\{i, j\}$, we have

$$a_i = \sum_{j=1}^n c_{i,j} > 0, \quad b_j = \sum_{i=1}^m c_{i,j} > 0.$$

If $c_{i,j} > 0$ for a pair of (i, j) , then $v_{i,j}(\mathbf{x})$ is a density. From Eq (2.6), we can rewrite each elliptical uniform u_j as

$$u_j(\mathbf{x}) = \sum_{i=1}^m \frac{c_{i,j}}{b_j} v_{i,j}(\mathbf{x}).$$

Following the argument presented in Theorem 2.1, we have

$$|\Sigma_{v_{i,j}}|^{\frac{1}{2}} \geq \frac{c_{i,j}}{b_j} |\Sigma_{u_j}|^{\frac{1}{2}},$$

with equality holding if and only if $v_{i,j}$ is elliptical uniform having “thickness” satisfying

$$\max\{c_{i,j} v_{i,j}(\mathbf{x})\} = \max\{b_j u_j(\mathbf{x})\}.$$

Similarly, rewriting each mixture component g_i in terms of $v_{i,j}$, we obtain

$$g_i(\mathbf{x}) = \sum_{j=1}^n \frac{c_{i,j}}{a_i} v_{i,j}(\mathbf{x}) \equiv \sum_{j=1}^n s_{i,j} v_{i,j}(\mathbf{x}).$$

Next, we create new *spherical unimodal* densities \tilde{g}_i 's corresponding to each g_i to facilitate lower boundings of $|\Sigma_{g_i}|$. Define \tilde{g}_i as follows:

$$\tilde{g}_i(\mathbf{x}) = \sum_{j=1}^n \frac{c_{i,j}}{a_i} \tilde{v}_{i,j}(\mathbf{x}) \equiv \sum_{j=1}^n s_{i,j} \tilde{v}_{i,j}(\mathbf{x}).$$

In the above, each $\{\tilde{v}_{i,1}, \dots, \tilde{v}_{i,n}\}$ are *spherical uniforms* whose means coincide and such that

$$\max\{c_{i,j} \tilde{v}_{i,j}(\mathbf{x})\} = \max\{b_j u_j(\mathbf{x})\}$$

for all $\{i, j\}$, hence yielding

$$|\Sigma_{\tilde{v}_{i,j}}|^{\frac{1}{2}} = \frac{c_{i,j}}{b_j} |\Sigma_{u_j}|^{\frac{1}{2}}.$$

Computing the determinant of the covariance matrix of g_i , we have

$$\begin{aligned} |\Sigma_{g_i}| &= |(s_{i,1} \Sigma_{v_{i,1}} + \dots + s_{i,n} \Sigma_{v_{i,n}}) + (s_{i,1} \mu_{v_{i,1}} \mu_{v_{i,1}}^T + \dots + s_{i,n} \mu_{v_{i,n}} \mu_{v_{i,n}}^T)| \\ &\geq |s_{i,1} \Sigma_{v_{i,1}} + \dots + s_{i,n} \Sigma_{v_{i,n}}| \\ &\geq (s_{i,1} |\Sigma_{v_{i,1}}|^{\frac{1}{d}} + \dots + s_{i,n} |\Sigma_{v_{i,n}}|^{\frac{1}{d}})^d \\ &\geq (s_{i,1} |\Sigma_{\tilde{v}_{i,1}}|^{\frac{1}{d}} + \dots + s_{i,n} |\Sigma_{\tilde{v}_{i,n}}|^{\frac{1}{d}})^d \\ &= |s_{i,1} \Sigma_{\tilde{v}_{i,1}} + \dots + s_{i,n} \Sigma_{\tilde{v}_{i,n}}| \\ &= |\Sigma_{\tilde{g}_i}|. \end{aligned}$$

The first inequality holds as a result of

$$(2.7) \quad |K_1 + K_2| \geq |K_1|,$$

where K_1 and K_2 are both non-negative definite symmetric $d \times d$ matrices. The second inequality holds because

$$(2.8) \quad |K_1 + K_2|^{\frac{1}{d}} \geq |K_1|^{\frac{1}{d}} + |K_2|^{\frac{1}{d}},$$

with identity holding if and only if K_1 and K_2 are proportional. The proof of both Eqs (2.7) and (2.8) can be found in Cover and Thomas (1988). The third inequality holds as we must have

$$|\Sigma_{v_{i,j}}| \geq |\Sigma_{\tilde{v}_{i,j}}|$$

as a direct result of Lemma 2.1. The equality that follows the third inequality is again a result of Eq (2.8), as all $\Sigma_{\tilde{v}_{i,j}}$'s are proportional to the identity matrix. We have just shown that

$$|\Sigma_{g_i}| \geq |\Sigma_{\tilde{g}_i}|$$

for all g_i , *i.e.* the pseudo-volume of each g_i is minimized when g_i is spherical unimodal. Therefore, a sufficient condition to (Claim 1) is

$$(\text{Claim 2}) \quad |\Sigma_f|^{\frac{1}{2}} \leq |\Sigma_{\tilde{g}_1}|^{\frac{1}{2}} + \dots + |\Sigma_{\tilde{g}_m}|^{\frac{1}{2}}.$$

Since f is elliptical unimodal, it is possible to find a corresponding spherical unimodal density f^s such that the hypervolumes are preserved, *i.e.* $|f^s| = |f|$. To prove (Claim 2), we only have to deal with the pseudo-volumes of spherical unimodal densities. We obtain $|\Sigma_{\tilde{g}_i}|$ as follows

$$|\Sigma_{\tilde{g}_i}| = \frac{1}{(d+2)^d} \cdot \left(\frac{c_{i,1}^{1+\frac{2}{d}} + \dots + c_{i,n}^{1+\frac{2}{d}}}{c_{i,1} + \dots + c_{i,n}} \right)^d.$$

Here, we make use of the fact that the covariance of a d -dimensional spherical uniform density defined by

$$u(\mathbf{x}) \propto \mathbb{I}_{|\mathbf{x}-\mu|<r}$$

is given as

$$\Sigma_u = \frac{r^2}{(d+2)} \cdot \mathbf{I}_d$$

where \mathbf{I}_d denotes the identity matrix in d -dimensional space. Refer to Eq (5.5). Similarly,

$$|\Sigma_f| = \frac{1}{(d+2)^d} \cdot \left(\frac{b_1^{1+\frac{2}{d}} + \dots + b_n^{1+\frac{2}{d}}}{b_1 + \dots + b_n} \right)^d.$$

Hence, proving (Claim 2) is equivalent to proving

(Claim 3)

$$\left(\frac{b_1^{1+\frac{2}{d}} + \dots + b_n^{1+\frac{2}{d}}}{b_1 + \dots + b_n} \right)^{\frac{d}{2}} \leq \left(\frac{c_{1,1}^{1+\frac{2}{d}} + \dots + c_{1,n}^{1+\frac{2}{d}}}{c_{1,1} + \dots + c_{1,n}} \right)^{\frac{d}{2}} + \dots + \left(\frac{c_{m,1}^{1+\frac{2}{d}} + \dots + c_{m,n}^{1+\frac{2}{d}}}{c_{m,1} + \dots + c_{m,n}} \right)^{\frac{d}{2}},$$

where $b_j = c_{1,j} + \dots + c_{m,j}$ for all j . To prove (Claim 3), we just have to invoke Lemma 2.2 given below for a total of $(m-1)$ times, adding up summands on the RHS two at a time and maintaining the “ \leq ” sign. We are now left with proof of Lemma 2.2 to prove Theorem 2.3.

Lemma 2.2. *Let a_i, b_i, c_i be sequences of non-negative real numbers such that for all i , $a_i = b_i + c_i$ and $a_i > 0$. Then the following inequality holds for any positive integers d and n .*

$$\left(\frac{a_1^{1+\frac{2}{d}} + \dots + a_n^{1+\frac{2}{d}}}{a_1 + \dots + a_n} \right)^{\frac{d}{2}} \leq \left(\frac{b_1^{1+\frac{2}{d}} + \dots + b_n^{1+\frac{2}{d}}}{b_1 + \dots + b_n} \right)^{\frac{d}{2}} + \left(\frac{c_1^{1+\frac{2}{d}} + \dots + c_n^{1+\frac{2}{d}}}{c_1 + \dots + c_n} \right)^{\frac{d}{2}}.$$

Equality holds if and only if the sequences a_i, b_i and c_i are linearly dependent.

Proof. The proof is similar to that of Chia and Nakano (2009), with the only difference being in d . We proceed in the spirit of Hardy *et al* (1988), as well as Pölya and Szegö (1972). Set $\mathbf{x} \equiv [x_1, \dots, x_n]^T$, $\mathbf{y} \equiv [y_1, \dots, y_n]^T$ and $\mathbf{z} \equiv$

$[z_1, \dots, z_n]^T$ and similarly for $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Let $\mathbf{x} = t\mathbf{y} + (1-t)\mathbf{z}$, *i.e.* $x_i = ty_i + (1-t)z_i$ for all i . Furthermore, define the function f as follows:

$$f(\mathbf{x}) = \left(\frac{x_1^{1+\frac{2}{d}} + \dots + x_n^{1+\frac{2}{d}}}{x_1 + \dots + x_n} \right)^{\frac{d}{2}},$$

and set

$$\phi(t) = f\{t\mathbf{y} + (1-t)\mathbf{z}\} \equiv f(\mathbf{x}),$$

where $0 \leq t \leq 1$. It suffices to prove that $\phi''(t) \geq 0$ for $0 \leq t \leq 1$. This is an immediate consequence of Jensen's inequality as $\phi''(t) \geq 0$ implies

$$\phi(t) \leq t\phi(0) + (1-t)\phi(1).$$

Setting $t = \frac{1}{2}$, we have

$$f\left(\frac{\mathbf{y}}{2} + \frac{\mathbf{z}}{2}\right) \leq \frac{1}{2}f(\mathbf{y}) + \frac{1}{2}f(\mathbf{z}).$$

Denoting by $\mathbf{y} = \mathbf{b}$, $\mathbf{z} = \mathbf{c}$, this becomes

$$f\left(\frac{\mathbf{a}}{2}\right) \leq \frac{1}{2}f(\mathbf{b}) + \frac{1}{2}f(\mathbf{c}).$$

However, from the definition of f , we must have

$$f\left(\frac{\mathbf{a}}{2}\right) = \frac{1}{2}f(\mathbf{a}).$$

Therefore $\phi''(t) \geq 0$ implies $f(\mathbf{a}) \leq f(\mathbf{b}) + f(\mathbf{c})$ as required. Equality holds if and only if $\phi''(t) = 0$.

We shall begin from the definition of ϕ as follows:

$$\phi(t) = f(\mathbf{x}) = (\sum x_i^{1+\frac{2}{d}})^{\frac{d}{2}} (\sum x_j)^{-\frac{d}{2}}.$$

Differentiating ϕ twice with respect to t and rearranging, we have

$$\begin{aligned} \frac{\phi''(t)}{\phi(t)} &= \frac{d(d+2)}{4} \cdot \underbrace{\left\{ \frac{\sum(y_k - z_k)}{\sum x_j} - \frac{\sum x_k^{\frac{2}{d}}(y_k - z_k)}{\sum x_i^{1+\frac{2}{d}}} \right\}^2}_A \\ &\quad + \left(\frac{d+2}{d} \right) \cdot (\sum x_i^{1+\frac{2}{d}})^{-2} \cdot \underbrace{\left[(\sum x_i^{1+\frac{2}{d}}) \cdot \left\{ \sum x_j^{\frac{2}{d}-1}(y_j - z_j)^2 \right\} - \left\{ \sum x_k^{\frac{2}{d}}(y_k - z_k) \right\}^2 \right]}_B. \end{aligned}$$

The term A is expressible as a square and therefore greater or equal to 0. To evaluate B , we set $p_i^2 = x_i^{1+\frac{2}{d}}$ and $q_j^2 = x_j^{\frac{2}{d}-1}(y_j - z_j)^2$, yielding

$$B = (\sum p_i^2) \cdot (\sum q_j^2) - (\sum p_k q_k)^2 \geq 0$$

via Cauchy-Schwarz's inequality. Therefore we must have

$$\phi''(t) \geq 0$$

due to the non-negativeness of x_i, y_i and z_i . Hence, Lemma 2.2, and consequently, Theorem 2.3 is proved. \square

As a result of Theorem 2.3, all elliptical unimodal densities with finite second moments are \mathcal{M} -undecomposable. Conversely, any density, which is \mathcal{M} -decomposable, cannot be elliptical unimodal. One can do better than that. In the next subsection, we further show that if f is \mathcal{M} -decomposable, then there exists an approximation to represent f via a mixture of Gaussian densities, which improves estimation of f .

2.5. Estimation of \mathcal{M} -Decomposable Densities.

Theorem 2.4 (Inequality on \mathcal{M} -Decomposable Densities). *Let f be probability density functions defined on $\mathbf{x} \in \mathcal{R}^d$. Let $\{g_1, \dots, g_m\}$ be a set of mixture components of f such that*

$$f(\mathbf{x}) = \alpha_1 g_1(\mathbf{x}) + \dots + \alpha_m g_m(\mathbf{x}),$$

where $0 < \alpha_j < 1$ and $\sum \alpha_j = 1$. Then the following result applies:

$$\begin{aligned} |\Sigma_f|^{\frac{1}{2}} &> |\Sigma_{g_1}|^{\frac{1}{2}} + \dots + |\Sigma_{g_m}|^{\frac{1}{2}} \\ \Rightarrow KL(f \parallel \tilde{f}) &> KL(f \parallel \alpha_1 \tilde{g}_1 + \dots + \alpha_m \tilde{g}_m). \end{aligned}$$

Here, $KL(p \parallel q)$ denotes the Kullback-Leibler divergence between densities p and q , given as

$$KL(p \parallel q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}.$$

Furthermore, \tilde{f} denotes the Gaussian density whose mean and covariance matrix coincide with those of f , and \tilde{g} 's are similarly defined.

Proof. We only need to prove that

$$(\text{Claim A}) \quad \int f(\mathbf{x}) \log \tilde{f}(\mathbf{x}) d\mathbf{x} < \int f(\mathbf{x}) \log \{\alpha_1 \tilde{g}_1(\mathbf{x}) + \dots + \alpha_m \tilde{g}_m(\mathbf{x})\} d\mathbf{x}.$$

Now, RHS of (Claim A)

$$\begin{aligned} &= \int \{\alpha_1 g_1(\mathbf{x}) + \dots + \alpha_m g_m(\mathbf{x})\} \cdot \log \{\alpha_1 \tilde{g}_1(\mathbf{x}) + \dots + \alpha_m \tilde{g}_m(\mathbf{x})\} d\mathbf{x} \\ &\geq \alpha_1 \int g_1(\mathbf{x}) \log \{\alpha_1 \tilde{g}_1(\mathbf{x})\} d\mathbf{x} + \dots + \alpha_m \int g_m(\mathbf{x}) \log \{\alpha_m \tilde{g}_m(\mathbf{x})\} d\mathbf{x} \\ &= \alpha_1 \{\log \alpha_1 + \int g_1(\mathbf{x}) \log \tilde{g}_1(\mathbf{x}) d\mathbf{x}\} + \dots + \alpha_m \{\log \alpha_m + \int g_m(\mathbf{x}) \log \tilde{g}_m(\mathbf{x}) d\mathbf{x}\}. \end{aligned}$$

From definitions, the probability density function of $\tilde{g}(\mathbf{x})$ is given by

$$\tilde{g}(\mathbf{x}) = (2\pi)^{-\frac{d}{2}} |\Sigma_g|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2} (\mathbf{x} - \mu_g)^T \Sigma_g^{-1} (\mathbf{x} - \mu_g)\right\},$$

where μ_g and Σ_g denote the mean and covariance matrix of g . We obtain

$$\int g(\mathbf{x}) \log \tilde{g}(\mathbf{x}) d\mathbf{x} = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_g| - \frac{d}{2}.$$

Hence, RHS of (Claim A)

$$\geq \alpha_1 \left\{ \log \alpha_1 - \frac{1}{2} \log |\Sigma_{g_1}| \right\} + \dots + \alpha_m \left\{ \log \alpha_m - \frac{1}{2} \log |\Sigma_{g_m}| \right\} - \frac{d}{2} \log(2\pi) - \frac{d}{2}.$$

Meanwhile,

$$\text{LHS of (Claim A)} = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_f| - \frac{d}{2}.$$

To complete the prove of Theorem 2.4, it suffices to demonstrate that

$$(\text{Claim B}) \quad \alpha_1 \log \frac{|\Sigma_{g_1}|^{\frac{1}{2}}}{\alpha_1} + \dots + \alpha_m \log \frac{|\Sigma_{g_m}|^{\frac{1}{2}}}{\alpha_m} < \log |\Sigma_f|^{\frac{1}{2}}.$$

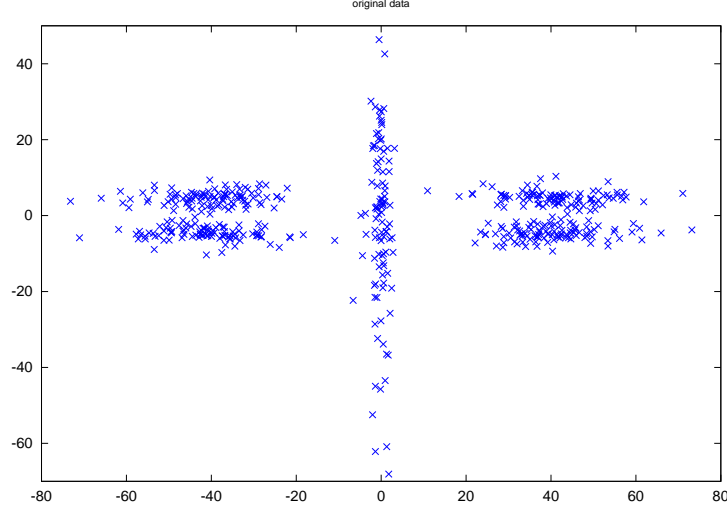


FIGURE 1. Original data from multimodal density drawn from mixture of five logistic densities.

Using Jensen's inequality, we have

$$\begin{aligned}
 \text{LHS of (Claim B)} &\leq \log\left(\alpha_1 \frac{|\Sigma_{g_1}|^{\frac{1}{2}}}{\alpha_1} + \dots + \alpha_m \frac{|\Sigma_{g_m}|^{\frac{1}{2}}}{\alpha_m}\right) \\
 &= \log(|\Sigma_{g_1}|^{\frac{1}{2}} + \dots + |\Sigma_{g_m}|^{\frac{1}{2}}) \\
 &< \log |\Sigma_f|^{\frac{1}{2}} = \text{RHS of (Claim B)},
 \end{aligned}$$

which completes the proof of Theorem 2.4 \square

We summarize the result of Theorem 2.4 as follows. Let f be any density in d -dimensional space. If f is \mathcal{M} -decomposable, then by definition, one can find a set of mixture components of f , such that the sum of pseudo-volumes of the mixture components is less than the pseudo-volume of the original density f . From Theorem 2.3, f cannot belong to the class of elliptical unimodal densities. It is possible to do better than that. Theorem 2.4 shows that f is better estimated via a weighted Gaussian mixture, rather than a single Gaussian density. The Gaussian components are created via moments matching of the mixture components of f . The better goodness of fit by the resultant weighted Gaussian mixture estimate is guaranteed in Kullback-Leibler sense. It should be noted that the analytical form of the original density f does not need to be known. In the next section, we demonstrate the use of Theorems 2.3 and 2.4 to statistical applications, namely cluster analysis and kernel density estimation.

3. APPLICATIONS USING \mathcal{M} -DECOMPOSABILITY

3.1. Clustering via \mathcal{M} -Decomposability: The Power of Two. One straightforward application of \mathcal{M} -decomposability is cluster analysis. Many existing clustering algorithms divide the dataset into clusters, based on the following heuristic: That the within-variances of clusters are minimized while the between-variance is

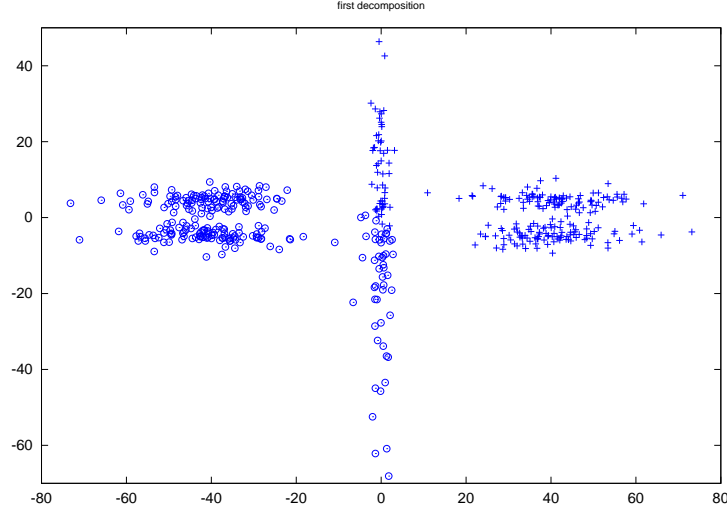


FIGURE 2. Original data split into two mixture components, represented by two different symbols. The sum of pseudo-volumes of the mixture components is less than that of original.

maximized at the same time. Another variation to this heuristic is to determine cluster allocations such that a function of volume of clusters is minimized. In particular, Shioda and Tunçel (2005) proposed dividing the dataset into k clusters, such that the total sum of MVE (minimum volume of ellipsoid) of k clusters are globally minimized. While the details for each algorithm may differ, the underlying idea is conceptually similar. Theorem 2.4 provides theoretical justification for minimizing sum of pseudo-volumes, and therefore supports all similar approaches of existing algorithms.

Intuitively, the rigorous approach to implement cluster analysis via Theorem 2.4 is to divide the dataset into $k(\geq 1)$ clusters, such that the sum of pseudo-volumes of all clusters are globally minimized. This approach is computationally unfeasible for dataset of any reasonable size. To this end, we propose the following alternative approach that captures the essence of Theorem 2.4 as far as possible. We devise a split-merge clustering strategy that involves splitting and merging, two clusters at a time. This lowers the overall computational load. We show that with our approach, the algorithm is able to overcome local minima. Consequently, it is possible to perform cluster analysis well, even with $k(> 2)$ clusters.

From the given sample $F = \{X_1, \dots, X_n\}$, we are interested to know if the original sample is \mathcal{M} -decomposable. We check if F can be partitioned into two clusters, such that the sum of pseudo-volumes of the clusters is less than that of F . We denote as $\{G, H\}$, a partition of F , such that

$$G = \{Y_1, \dots, Y_m\}, \quad H = \{Y_{m+1}, \dots, Y_n\}$$

and $G \cup H = F$, with Y 's being a rearrangement of X . We further denote the sample covariance matrices of F, G, H as $\mathcal{S}_F, \mathcal{S}_G$ and \mathcal{S}_H . Our task is to find the

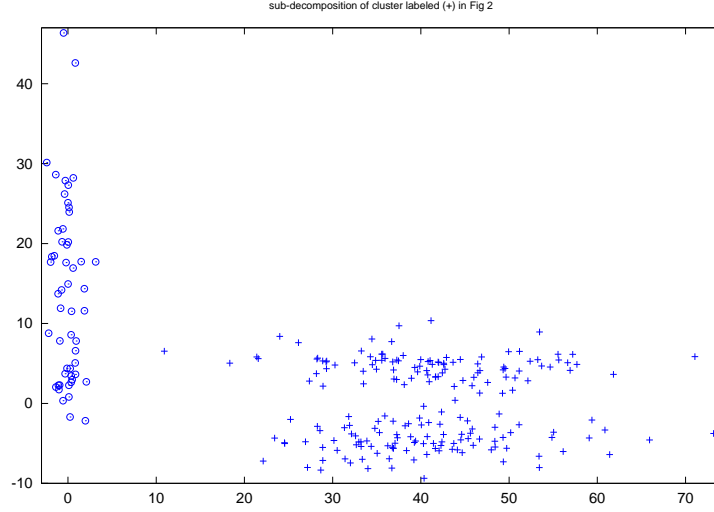


FIGURE 3. Mixture component denoted by (+) in Fig 2 split into two further mixture components.

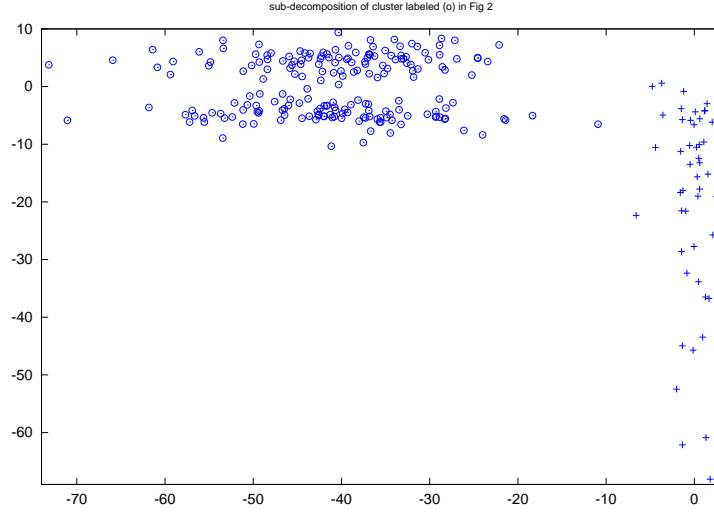


FIGURE 4. Mixture component denoted by (o) in Fig 2 split into two further mixture components.

optimal partition $\{G, H\}$ such that

$$|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}}$$

is globally minimized and test this value against $|\mathcal{S}_F|^{\frac{1}{2}}$. If

$$(3.1) \quad \frac{|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}}}{|\mathcal{S}_F|^{\frac{1}{2}}} < 1 + \tau_s,$$

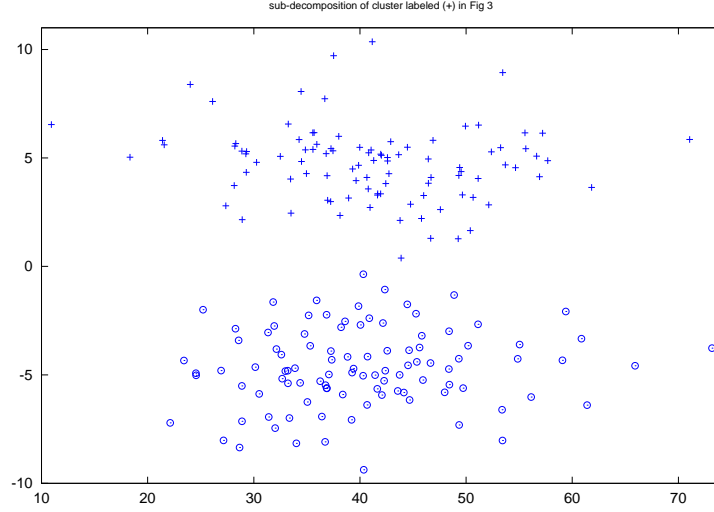


FIGURE 5. Mixture component denoted by (+) in Fig 3 split into two further mixture components.

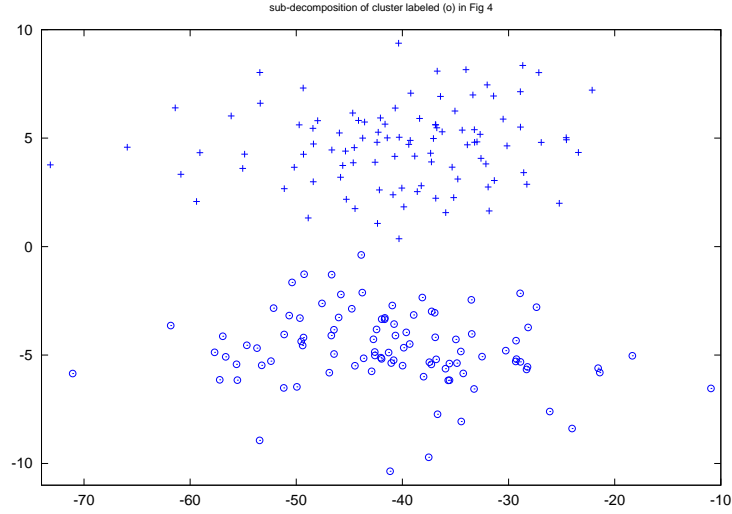


FIGURE 6. Mixture component denoted by (o) in Fig 4 split into two further mixture components.

where τ_s is a threshold value close to zero, then, we can conclude that F is likely to be \mathcal{M} -decomposable. However, if Eq (3.1) is not satisfied, then F is likely to be \mathcal{M} -undecomposable. To robustify the “splitting process” against local minima traps, it is possible to set the RHS of Eq (3.1) to be greater than 1. Furthermore, taking into consideration error due to finiteness of sample sizes, imperfection of splitting

algorithms, and also accounting for limiting the number of mixture components to two, we recommend that the τ_s on the RHS of Eq (3.1) to be about 0.05.

When one concludes that a particular cluster F is probably \mathcal{M} -undecomposable, it is possible to stop at one cluster. However, if F is found to be \mathcal{M} -decomposable into clusters of G and H , one may repeat the splitting process for G and H . The process is then reiterated until all clusters are probably \mathcal{M} -undecomposable. When that happens, the splitting process ends.

Our strategy also includes “merging” of clusters. At the point when all splitted clusters are probably \mathcal{M} -undecomposable, we select two clusters at a time and perform the following test. Now, let Q, R denote the two chosen clusters and P be the union of the two clusters, *i.e.* $P = Q \cup R$. We then check the sum of the pseudo-volumes of Q and R and compare against that of P . If

$$(3.2) \quad \frac{|\mathcal{S}_Q|^{\frac{1}{2}} + |\mathcal{S}_R|^{\frac{1}{2}}}{|\mathcal{S}_P|^{\frac{1}{2}}} \geq 1 + \tau_m,$$

we conclude that Q and R should be merged to form a larger cluster P . This process is repeated until there are no more mergeable clusters left. To prevent overclustering, we recommend τ_m to be around -0.05 .

We have described a possible algorithm using \mathcal{M} -decomposability to perform cluster analysis. The crucial point is to find a partition $\{G, H\}$ such that $|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}}$ is minimized as far as possible. There are many possible approaches to this task. To find the global minimum of the sum $|\mathcal{S}_G|^{\frac{1}{2}} + |\mathcal{S}_H|^{\frac{1}{2}}$ is computationally unfeasible and may be NP-hard. Here, we propose a computationally simpler approach. At each spitting step, we simply fit a two-mixture Gaussian to the original cluster F , and then run the EM algorithm to convergence to obtain the partition $\{G, H\}$. However, we emphasize that the EM algorithm approach itself is not critical, and that it is possible to use other approaches to obtain a reasonable partition $\{G, H\}$ of F at the splitting step. The main point here is the concept of clustering via \mathcal{M} -decomposability. In the two examples presented below, we show that it is possible to perform clustering analysis reasonably well, using our proposed algorithm.

3.2. Clustering of Simulated Data. The simulation example provided here is drawn from a five-mixture logistic densities as follows. The sample F is generated by 100 samples each from five logistic densities with the following means and covariance matrices:

$$\begin{aligned} L_1 &: \left[\begin{pmatrix} -40 \\ 5 \end{pmatrix}, \begin{pmatrix} 12\pi^2 & 0 \\ 0 & \frac{\pi^2}{3} \end{pmatrix} \right] \\ L_2 &: \left[\begin{pmatrix} -40 \\ -5 \end{pmatrix}, \begin{pmatrix} 12\pi^2 & 0 \\ 0 & \frac{\pi^2}{3} \end{pmatrix} \right] \\ L_3 &: \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\pi^2}{3} & 0 \\ 0 & 48\pi^2 \end{pmatrix} \right] \\ L_4 &: \left[\begin{pmatrix} 40 \\ 5 \end{pmatrix}, \begin{pmatrix} 12\pi^2 & 0 \\ 0 & \frac{\pi^2}{3} \end{pmatrix} \right] \\ L_5 &: \left[\begin{pmatrix} 40 \\ -5 \end{pmatrix}, \begin{pmatrix} 12\pi^2 & 0 \\ 0 & \frac{\pi^2}{3} \end{pmatrix} \right] \end{aligned}$$

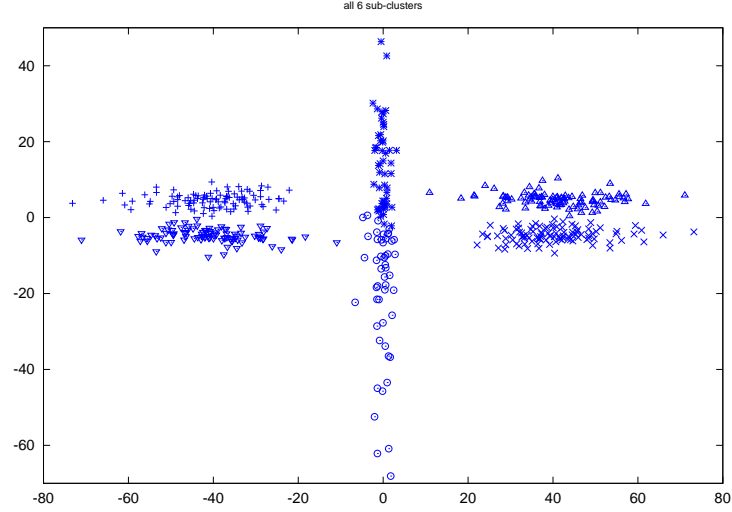


FIGURE 7. All six \mathcal{M} -undecomposable clusters of original data, represented by six different symbols.

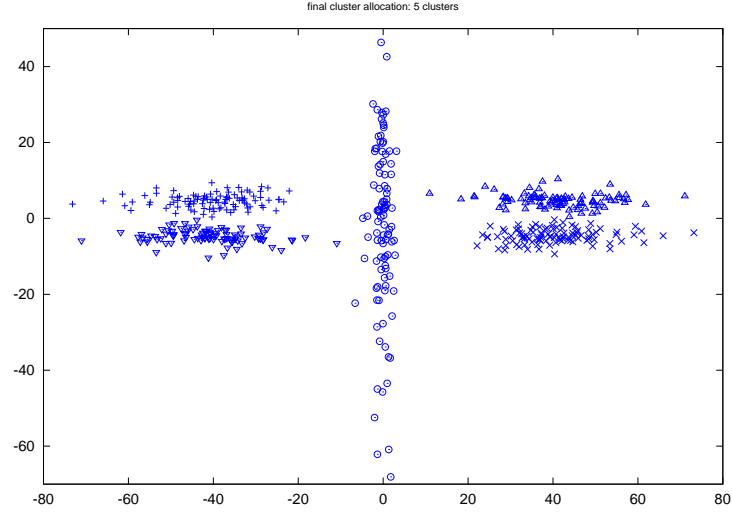


FIGURE 8. Final cluster allocation formed by merging clusters from Fig 7. Five clusters are recovered faithfully.

Fig 1 shows the original sample F . Clustering is performed without knowledge of either the number of clusters or the functional form of the clusters. At the first split step, we fit a two-Gaussian mixture to F , and perform EM to obtain the partition $\{G, H\}$. The result is shown in Fig 2. As Eq (3.1) is satisfied for F, G, H , we split F into G and H . This is a case of EM converging to a local minima as it is (visually) unlikely that G and H are meaningful clusters of F . However, from Eq (3.1), it

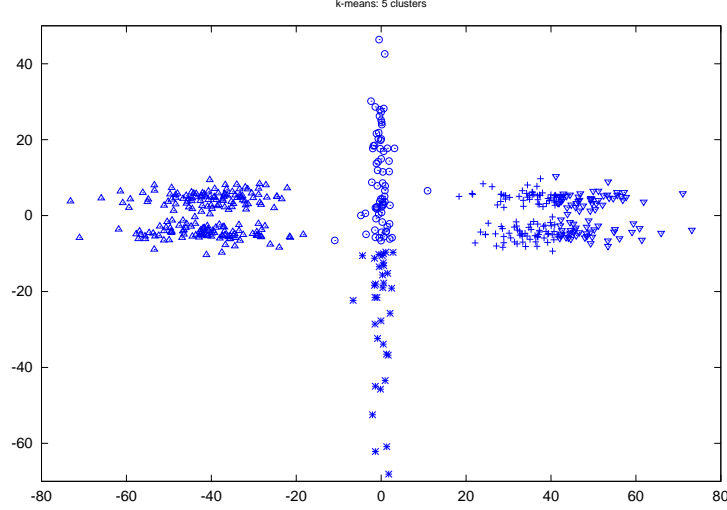


FIGURE 9. Final cluster allocation via k -means, represented by five different symbols. k -means algorithm fails to recover clusters faithfully.

is theoretically better off to split F into G and H . The theoretical justification is given in Kullback-Leibler sense. The splitting process is repeated for G and H and the results are shown in Figs 3 and 4. The splitting process continues until we arrive at six clusters that are all \mathcal{M} -undecomposable (Fig 7). Finally, we begin the merging process and find that the two clusters Q , shown as asterisk (*) and R , shown as circle (o) in Fig 7, satisfy Eq (3.1) where $P = Q \cup R$. The two clusters are then merged and we are left with five clusters shown in Fig 8. This example shows that our algorithm is easy to implement and is robust to local minima.

A popular clustering algorithm is the k -means method, which is optimal for nearly spherical clusters. However, it does not work here because of the presence of inherently elongated clusters. Even by setting $k = 5$, the k -means method does not achieve a meaningful cluster allocation, as shown in Fig 9. Cluster analysis via k -means is sensitive to rescaling of axes, because k -means involves comparison of distances. To improve the performance of k -means analysis, there exist many pre-processing heuristics, *e.g.* rescaling the axes such that all axial units or marginal standard deviations become compatible. For this simulation example, rescaling is unlikely to improve cluster analysis via k -means because elongated clusters are not likely to be eliminated. On the other hand, cluster analysis via \mathcal{M} -decomposability involves comparison of pseudo-volumes instead of distances, and are therefore *invariant* to rescaling of axes.

3.3. Clustering of Iris Dataset. Next, we analyze Fisher’s Iris dataset via \mathcal{M} -decomposability. The dataset was obtained from Asuncion and Newman (2007). The dataset consists of 150 four-dimensional data. The four attribute information given are sepal length, sepal width, petal length and petal width, all in centimetres. There are altogether three classes, namely “Setosa”, “Versicolor” and “Virginica”, in the proportion of 50 : 50 : 50.

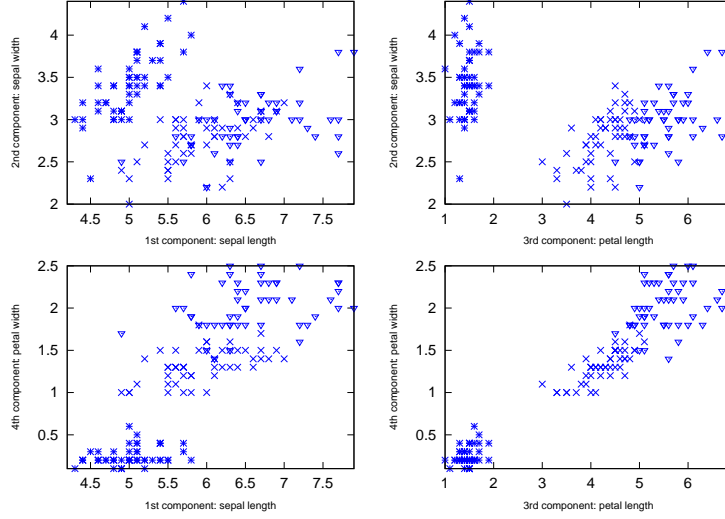


FIGURE 10. True Iris data: setosa(asterix), versicolor(cross), virginica(triangle).

We perform cluster analysis of the dataset via \mathcal{M} -decomposability, without knowledge of the actual number of classes. At the end of the analysis, we confirm that there are altogether three classes, in the proportion of 50 : 45 : 55. The first 50 data coincide with “Setosa” (0 misspecification). For “Versicolor” and “Virginica”, there are altogether five misspecifications. (Five “Versicolor” are mislabeled as “Virginica”). The data is depicted graphically in Fig 10 (true class) and Fig 11 (estimated class).

Although our analysis results in five cases of misspecifications, our allocation of “Versicolor” and “Virginica” achieves a smaller pseudo-volume than the “true class”. Denoting the “true” classification of “Versicolor” and “Virginica” by $\{v_1, v_2\}$, and our estimation by $\{\hat{v}_1, \hat{v}_2\}$ respectively, our estimation yields

$$|\Sigma_{\hat{v}_1}|^{\frac{1}{2}} + |\Sigma_{\hat{v}_2}|^{\frac{1}{2}} \approx 0.01563,$$

as compared to

$$|\Sigma_{v_1}|^{\frac{1}{2}} + |\Sigma_{v_2}|^{\frac{1}{2}} \approx 0.01587.$$

The pseudo-volume of “Versicolor” and “Virginica” combined into a single class is approximately 0.01799.

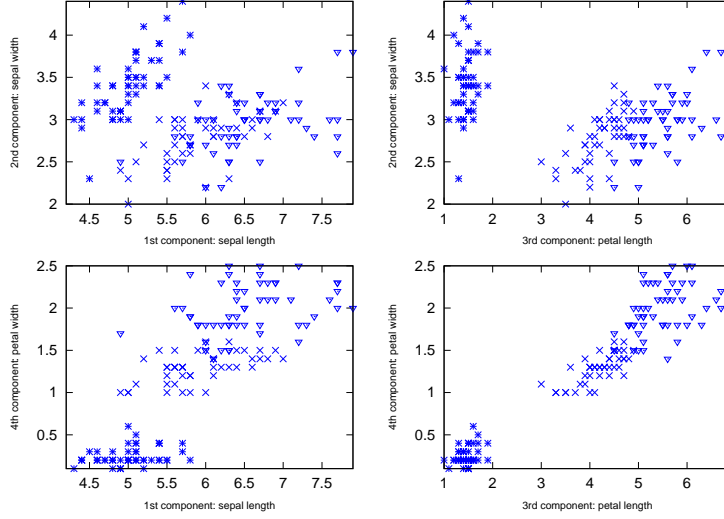


FIGURE 11. Iris data recovered via \mathcal{M} -decomposability: setosa(asterix), versicolor(cross), virginica(triangle).

3.4. Kernel Density Estimation. Density estimation is an important statistical tool that is widely used in many scientific and engineering fields. Given raw measurements or data, the task is to recover the unknown density from which the original data is generated. The problem statement is as follows. Given $\{X_1, \dots, X_n\}$, which is generated from an unknown distribution with density f , the task is to estimate f . For simplicity, we consider only univariate density estimation.

In density estimation, it is usually difficult to determine quantitatively the number of modes in the underlying distribution, just from the given data. In this respect, Theorem 2.4 can be used for parametric density estimation via Gaussian mixtures. Besides via Gaussian mixtures, a popular approach to density estimation is via the kernel density estimator. The kernel density estimator approach is non-parametric and is treated in detail in Scott (1992), Silverman (1986), Wand and Jones (1995), Härdle *et al* (2004). The formula for the kernel density estimator, given data $\{X_1, \dots, X_n\}$ is

$$(3.3) \quad \hat{f}(x; b) = (nb)^{-1} \sum_{i=1}^n K\{(x - X_i)/b\},$$

see, *e.g.* Wand and Jones (1995). Usually K is chosen to be a unimodal density that is symmetric about zero, and is called the *kernel*. The positive number b is called the *bandwidth*. Such a formulation ensures that $\hat{f}(x; b)$ is also a density. One property of the kernel density estimator is that the choice bandwidth is more important than the choice of the kernel itself. The optimal choice of the bandwidth ensures that the density estimate becomes optimally smoothed. One popular choice of the bandwidth is

$$(3.4) \quad b = n^{-\frac{1}{5}} \hat{\sigma},$$

where $\hat{\sigma}$ is the sample standard deviation of the given data and n denotes the sample size. One known problem of the bandwidth given in Eq (3.4) is that it works well for densities that are approximately symmetric unimodal. For multimodal densities, the bandwidth tends to produce an oversmoothed density.

Here, we propose an \mathcal{M} -decomposability based algorithm to improve kernel density estimation. As we are only dealing with the univariate case, we consider just the sorted data $F = \{X_{[1]}, \dots, X_{[n]}\}$. Similar to Section 3.1, we perform clustering of F via splitting and merging. In one-dimension, the splitting process becomes much simpler as we just have to find m ($2 < m < n - 1$) such that $(\sigma_G + \sigma_H)$ is minimized.

For clarity of explanation, we assume that the original data F has two clusters, and that $G = \{X_{[1]}, \dots, X_{[m]}\}$ and $H = \{X_{[m+1]}, \dots, X_{[n]}\}$ are the optimal partition of F . We also have $\sigma_G + \sigma_H < \sigma_F$. As such, we can expect the density estimation via the weighted mixture of G and H to be better than that of the original data set. Therefore, one may propose a mixture kernel density estimator \hat{f}_1 of F given as follows:

$$\hat{f}_1(x) = \frac{m}{n} \hat{g}(x; b_g) + \frac{n-m}{n} \hat{h}(x; b_h),$$

where

$$b_g = m^{-\frac{1}{5}} \hat{\sigma}_G, \quad b_h = (n-m)^{-\frac{1}{5}} \hat{\sigma}_H,$$

and

$$\begin{aligned} \hat{g}(x; b_g) &= (mb_g)^{-1} \sum_{i=1}^m K\{(x - X_{[i]})/b_g\}, \\ \hat{h}(x; b_h) &= \{(n-m)b_h\}^{-1} \sum_{i=m+1}^n K\{(x - X_{[i]})/b_h\}. \end{aligned}$$

The original kernel density estimator \hat{f} of F is given in Eq (3.3).

As an experiment, we generate a sample of size 1000 from a bimodal density, with functional form given as

$$f(x) = \frac{0.2}{\cosh^2(x + 2.5)} + \frac{0.3}{\cosh^2(x - 2.5)}.$$

The “true” density is shown as solid line in Figs 12, 13. By simply computing one single bandwidth b on the whole sample set, we obtain a kernel density estimator (computed using \hat{f}). The result is shown as crosses in Fig 13. By using \mathcal{M} -decomposability and splitting the data into two clusters, we obtain a mixture kernel density estimator (computed using \hat{f}_1). The result is shown as crosses in Fig 12. From Figs 12 and 13, it is clear that the kernel density estimator computed using \mathcal{M} -decomposability is closer to the true density. In this example, we see a pronounced effect of oversmoothing (Fig 13) for the kernel density estimator with a single bandwidth. This is because the original density is bimodal with modes well separated. The undesirable effect of oversmoothing is alleviated by implementing \mathcal{M} -decomposability.

4. CONCLUSION

In this paper, we generalized the notion of \mathcal{M} -decomposability proposed by Chia and Nakano (2009) to d -dimensions, where $d \geq 1$. Furthermore, we also

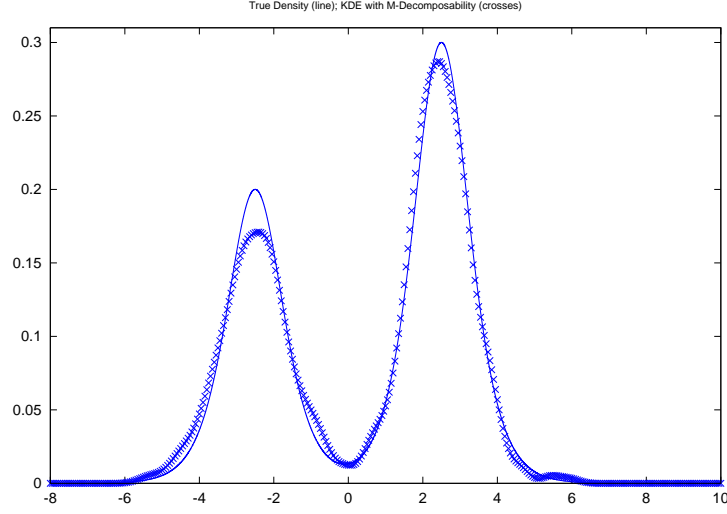


FIGURE 12. True density shown as line; kernel estimate with \mathcal{M} -decomposability shown as crosses.

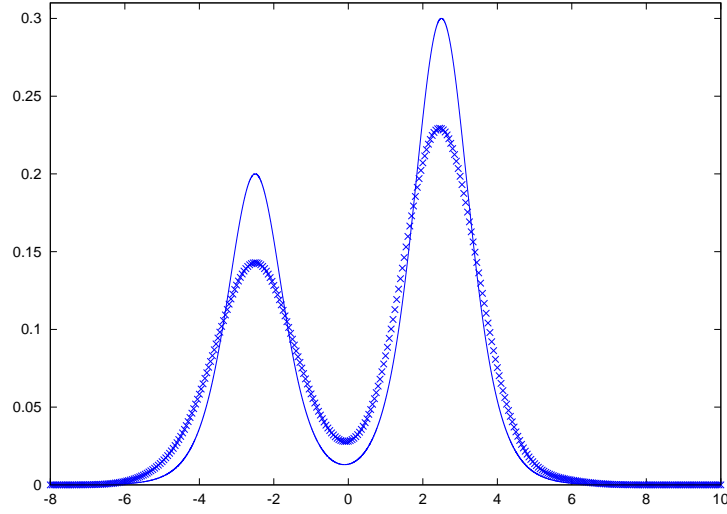


FIGURE 13. True density shown as line; kernel estimate with single bandwidth shown as crosses. Comparing with Fig 12, we see that for a multimodal density, \mathcal{M} -decomposability improves kernel density estimation.

broadened the scope of definition of \mathcal{M} -decomposability to accommodate any number of mixture components. We also derived two theorems pertaining to \mathcal{M} -decomposability. As a result of the first theorem, all elliptical unimodal densities

are \mathcal{M} -undecomposable. Consequently, any density that is \mathcal{M} -decomposable cannot belong to the class of elliptical unimodal densities, which includes many general densities, such as Gaussian, Laplace, uniform, logistic, *etc.* The second theorem goes further to say that if a density is \mathcal{M} -decomposable, then it is possible to model the density better via a weighted mixture of Gaussian densities. The goodness of fit here is defined in Kullback-Leibler sense. \mathcal{M} -decomposability is closely related to the modality of probability density functions, and hence the theoretical results derived from this paper should appeal to theoreticians and practitioners alike.

We proposed \mathcal{M} -decomposability as a criterion to determine the modality of a given density, *i.e.* if the density is unimodal or multimodal. A practical application is non-parametric cluster analysis. Here, one does not need to know the parametric model for the underlying clusters. The only assumption required is that the underlying clusters are approximately elliptical and unimodal. In this sense, clustering via \mathcal{M} -decomposability is more flexible and robust than clustering via parametric models or via k -means. Furthermore, we designed a clustering algorithm which automatically determines the number of clusters. Our algorithm have been tested on non-Gaussian cluster examples, as well as the popular Iris dataset. Another example of application of \mathcal{M} -decomposability is density estimation. We also devised a scheme to improve kernel density estimation.

Cluster analysis and kernel density estimation are closely related to statistical learning. Examples are given in Hastie *et al* (2001). Therefore, \mathcal{M} -decomposability will also be useful in areas such as independent component analysis [Comon (1994), Hyvärinen and Oja (2000)], machine learning [Hand *et al* (2001)], *etc.* Furthermore, as \mathcal{M} -decomposability has been demonstrated to improve density estimation, it may also be applied to the improvement of proposal densities in *Markov chain Monte Carlo* (MCMC) methodologies [Robert and Casella (2004)] and particle filtering. For example, in Kotecha and Djurić (2003), a class of particle filters, called Gaussian particle filters were introduced. To represent the prior density at each time-step, the authors generated particles from the Gaussian density fitted to the weighted particles representing the previous posterior density. Using Theorem 2.4, the estimation of the prior density can be improved by fitting a mixture of Gaussian densities to the weighted particles if necessary, using \mathcal{M} -decomposability as the criterion to determine the fit. Similarly, in Lee and Chia (2002), the authors used Gaussian densities as proposal densities to generate the next prior density via MCMC. Using \mathcal{M} -decomposability, it is possible to improve the proposal densities, which in turn enhances mixing and improves the acceptance rates of the sequential MCMC steps.

5. APPENDIX

5.1. Special Orthogonal Matrices. A class of matrices in d -dimensional space satisfying

$$A^{-1} = A^T, \quad |A| = 1$$

is given the name *special orthogonal matrices*, and denoted as $\mathcal{SO}(d)$. Special orthogonal matrices include all rotation matrices in d -dimensional space. They play an important role in the proof of Lemma 2.1. The next theorem, which is related to the representation of special orthogonal matrices, is brought to our attention from Bernstein (2005).

Theorem 5.1. *Let $A \in \mathcal{R}^{d \times d}$, where $d \geq 2$. Then $A \in \mathcal{SO}(d)$ if and only if there exist m such that $1 \leq m \leq d(d-1)/2$, $\theta_1, \dots, \theta_m \in \mathcal{R}$, and $j_1, \dots, j_m, k_1, \dots, k_m \in \{1, \dots, d\}$ such that*

$$A = \prod_{i=1}^m P(\theta_i, j_i, k_i),$$

where

$$P(\theta, j, k) \equiv \mathbf{I}_d + (\cos \theta - 1)(E_{j,j} + E_{k,k}) + (\sin \theta)(E_{j,k} - E_{k,j}).$$

Here, \mathbf{I}_d denotes the d -dimensional identity matrix and $E_{i,j}$ denotes the $d \times d$ matrix with one at the (i, j) -th element and zeros everywhere else.

The proof is given in Farebrother and Wrobel (2002).

Remark $P(\theta, j, k)$ is a *plane* or *Givens rotation*.

Remark Theorem 5.1 is an extension of Euler's rotation theorem, which is the case when $n = 3$.

5.2. Proof of Lemma 2.1. Without loss of generality, we set the mean of f to the origin to simplify computations. Next, note that it is possible to apply a linear transformation to the support space of f , such that the transformed density f^w satisfies

$$\Sigma_{f^w} = k_f \mathbf{I}_d, \quad |\Sigma_{f^w}| = |\Sigma_f|,$$

where \mathbf{I}_d denotes the d -dimensional identity matrix. As a result of the linear transformation, we must also have

$$\max(f^w) = \max(f).$$

Next, we denote by u the density of the spherical uniform that satisfies $\max(u) = \max(f^w)$. Our goal is then to prove that $|\Sigma_{f^w}| \geq |\Sigma_u|$, with identity holding if and only if $f^w = u$. In order to facilitate comparisons of pseudo-volumes of f^w and u , we shall construct a *spherical* density f^s (see Definition 2.2) that satisfies

$$\Sigma_{f^s} = \Sigma_{f^w}.$$

By construction, we have $|\Sigma_{f^s}| = |\Sigma_{f^w}|$ and therefore, an equivalent statement of our goal is $|\Sigma_{f^s}| \geq |\Sigma_u|$. The steps for the construction of f^s are given in the following paragraph.

We denote by f_j the resultant probability density function when a rotation operator $R_j \in \mathcal{SO}(d)$ is applied onto the support space of f^w . We have

$$\Sigma_{f_j} = \Sigma_{f^w} \quad \text{and} \quad \mu_{f_j} = \mathbf{0} = \mu_f.$$

In other words, the mean and covariance of f^w are *invariant to rotation* if $\Sigma_{f^w} \propto \mathbf{I}_d$. For any rotation operators $R_i, R_j \in \mathcal{SO}(d)$, any *weighted mixture* of f_i and f_j will again have the same mean and covariance matrix. Denoting the mixture by g , we have

$$g(\mathbf{x}) = \alpha f_i(\mathbf{x}) + (1 - \alpha) f_j(\mathbf{x}),$$

where $0 < \alpha < 1$. The covariance of g is given by

$$\Sigma_g = \alpha \Sigma_{f_i} + (1 - \alpha) \Sigma_{f_j} + \alpha(1 - \alpha)(\mu_{f_i} - \mu_{f_j})(\mu_{f_i} - \mu_{f_j})^T = \Sigma_{f^w}.$$

In two-dimensional space, a rotation operator can be represented as

$$R^\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

From Theorem 5.1, it is possible to represent any rotation in d -dimensional space as a product of Given's rotations shown below.

$$R = R_1^{\theta_1} \cdots R_D^{\theta_D}$$

where $D = d(d-1)/2$. We are ready to construct f^s as follows:

$$(5.1) \quad f^s(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^D \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} f^w(R_1^{\theta_1} \cdots R_D^{\theta_D} \mathbf{x}) d\theta_1 \cdots d\theta_D.$$

By construction, f^s is the uniform mixture of all possible rotations of the probability density function f^w in d -dimensional space. To show that $\Sigma_{f^s} = \Sigma_{f^w}$, note that

$$\begin{aligned} \Sigma_{f^s} &= \int \mathbf{x} \mathbf{x}^T f^s(\mathbf{x}) d\mathbf{x} \\ &= \left(\frac{1}{2\pi}\right)^D \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} \underbrace{\left\{ \int \mathbf{x} \mathbf{x}^T f^w(R_1^{\theta_1} \cdots R_D^{\theta_D} \mathbf{x}) d\mathbf{x} \right\}}_A d\theta_1 \cdots d\theta_D. \end{aligned}$$

The term A is simply the covariance matrix of the transformed density after applying rotation operator $R_1^{\theta_1} \cdots R_D^{\theta_D}$ to the support space of f^w . As Σ_{f^w} is invariant to rotation, we have

$$\Sigma_{f^s} = \left(\frac{1}{2\pi}\right)^D \underbrace{\left\{ \int_0^{2\pi} \cdots \int_0^{2\pi} d\theta_1 \cdots d\theta_D \right\}}_{D \text{ times}} \Sigma_{f^w} = \Sigma_{f^w}.$$

Furthermore, f^s must be *spherical* as one can easily verify that $f^s(R\mathbf{x}) = f^s(\mathbf{x})$ for any $R \in \mathcal{SO}(d)$. On top of these, from Eq (5.1), we have

$$(5.2) \quad f^s(\mathbf{x}) \leq \left(\frac{1}{2\pi}\right)^D \underbrace{\int_0^{2\pi} \cdots \int_0^{2\pi}}_{D \text{ times}} \max(f^w) d\theta_1 \cdots d\theta_D = \max(f^w).$$

We have therefore constructed a spherical density f^s whose covariance matrix is the same as that of f^w . Now we are left with proving that $|\Sigma_{f^s}| \geq |\Sigma_u|$ to complete the proof of the lemma.

We express the covariance matrix of u by $\Sigma_u = k_u \mathbf{I}_d$. Our goal will be accomplished if we can prove that $k_f \geq k_u$. From Eq (5.2), we have $f^s(\mathbf{x}) \leq \max(f) = M_f$, and the followings are straightforward:

- (1) $u(\mathbf{x}) \geq f^s(\mathbf{x})$ for $|\mathbf{x}| \leq R$, where $u(\mathbf{x}) = M_f$ throughout.
- (2) $u(\mathbf{x}) \leq f^s(\mathbf{x})$ for $|\mathbf{x}| > R$, where $u(\mathbf{x}) = 0$ throughout.

Here, R represents the radius of the spherical uniform u . Moreover, as f^s and u are both spherical and have means centred at the origin, there exist functions \tilde{f}^s and \tilde{u} such that

$$f^s(\mathbf{x}) = \tilde{f}^s(|\mathbf{x}|) = \tilde{f}^s(r); \quad u(\mathbf{x}) = \tilde{u}(|\mathbf{x}|) = \tilde{u}(r),$$

using Definition 2.2 and representation in the hyperspherical coordinates. Furthermore, we define $h(\mathbf{x}) \equiv f^s(\mathbf{x}) - u(\mathbf{x})$. Note that $h(\mathbf{x})$ is *not* a probability density function as $h(\mathbf{x})$ takes negative values and

$$(5.3) \quad \int h(\mathbf{x}) d\mathbf{x} = 0.$$

Using the hyperspherical coordinate representation, there must exist a function \tilde{h} such that $h(|\mathbf{x}|) = \tilde{h}(r)$, and

$$\tilde{h}(r) \begin{cases} \leq 0 & \text{for } r \leq R; \\ \geq 0 & \text{for } r > R. \end{cases}$$

Note that \tilde{h} is identically 0 if and only if $f^s = u$, or equivalently, f is elliptical uniform. Now,

$$\begin{aligned} k_f - k_u &= \mathbf{e}_1^T (\Sigma_{f^s} - \Sigma_u) \mathbf{e}_1 \\ &= \int \mathbf{e}_1^T \mathbf{x} \mathbf{x}^T \mathbf{e}_1 \{f^s(\mathbf{x}) - u(\mathbf{x})\} d\mathbf{x} \\ &= \int |\mathbf{e}_1^T \mathbf{x}|^2 h(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Here, \mathbf{e}_1 is the unit vector parallel to the first axis. Representation via spherical coordinates yields

$$\begin{aligned} k_f - k_u &= \int \cdots \int x_1^2 \tilde{h}(r) r^{d-1} \sin^{d-2}(\phi_1) \cdots \sin(\phi_{d-2}) dr d\phi_1 \cdots d\phi_{d-1} \\ &= \int_0^\infty r^{d+1} \tilde{h}(r) dr \times \Phi_1 \times \cdots \times \Phi_{d-1}, \end{aligned}$$

with

$$\Phi_1 = \int_0^\pi \cos^2(\phi_1) \sin^{d-2}(\phi_1) d\phi_1, \quad \Phi_{d-1} = 2\pi,$$

and the rest of Φ_i 's ($2 \leq i \leq d-2$) satisfying

$$\Phi_i = \int_0^\pi \sin^{d-i-1}(\phi_i) d\phi_i.$$

Apparently, all Φ_i 's are strictly positive and we only need to prove

$$(*) \quad \int_0^\infty r^{d+1} \tilde{h}(r) dr \geq 0$$

to arrive at the conclusion that $k_f \geq k_u$. Representing Eq (5.3) via hyperspherical coordinates, we have

$$\int_0^\infty r^{d-1} \tilde{h}(r) dr \times \int_0^\pi \sin^{d-2}(\phi_1) d\phi_1 \times \Phi_2 \times \cdots \times \Phi_{d-1} = 0$$

and therefore

$$\int_0^\infty r^{d-1} \tilde{h}(r) dr = 0.$$

To prove (*), we break up the integral into as follows:

$$\begin{aligned} \int_0^\infty r^{d+1} \tilde{h}(r) dr &= \int_0^R r^{d-1} r^2 \underbrace{\tilde{h}(r)}_{\leq 0} dr + \int_R^\infty r^{d-1} r^2 \underbrace{\tilde{h}(r)}_{\geq 0} dr \\ &\geq \int_0^R r^{d-1} R^2 \tilde{h}(r) dr + \int_R^\infty r^{d-1} R^2 \tilde{h}(r) dr \\ &= R^2 \times \int_0^\infty r^{d-1} \tilde{h}(r) dr = 0. \end{aligned}$$

Equality holds if and only if $\tilde{h} = 0$ identically, implying that f^s is spherical uniform, or in other words, f is elliptical uniform. This proves $k_f \geq k_u$ and consequently, we have

$$|\Sigma_f| = |\Sigma_{f^s}| \geq |\Sigma_u|.$$

Finally, we need to show that the pseudo-volume of an elliptical uniform density u with $\max(u) = M_u$ is given by

$$(5.4) \quad |\Sigma_u|^{\frac{1}{2}} = \frac{\Gamma(\frac{d}{2} + 1)}{M_u \{\pi(d+2)\}^{\frac{d}{2}}}.$$

We first compute the covariance matrix of an uniform density on the *hypersurface* of the d -dimensional sphere. Consider the probability mass function of a discrete random variable X given below:

$$f_X(\mathbf{x}) = \begin{cases} \frac{1}{2d} & \text{if } \mathbf{x} = \pm a \mathbf{e}_j, j = 1, \dots, d, \\ 0 & \text{otherwise.} \end{cases}$$

The covariance matrix of the above distribution is computed as $a^2 \mathbf{I}_d/d$. It is possible to generate an uniform density on the hypersurface of the d -dimensional sphere of radius a , by applying rotations to the discrete random variable given in Eq (5.4). Therefore, the covariance matrix of an uniform density on the hypersurface of a d -dimensional sphere of radius r is $a^2 \mathbf{I}_d/d$. By considering a spherical uniform density as a continuous mixture of hypersurfaces, we obtain the covariance matrix of a spherical uniform density with radius r as

$$(5.5) \quad \Sigma_u = \frac{\mathbf{I}_d}{d} \frac{\int_0^r a^{d+1} da}{\int_0^r a^{d-1} da} = \frac{r^2}{d+2} \mathbf{I}_d.$$

We therefore obtain the pseudo-volume of a spherical uniform density radius r as

$$|\Sigma_u|^{\frac{1}{2}} = \frac{r^d}{(d+2)^{\frac{d}{2}}}.$$

Using the fact that the volume of a d -dimensional sphere of radius r is given by

$$V = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(\frac{d}{2} + 1)}$$

we obtain the require pseudo-volume. Hence, the proof for Lemma 2.1 is complete. \square

5.3. Proof of Theorem 2.2. We can define the following continuous function on non-negative values of y , for a given f :

$$q(y) = \int \min\{f(\mathbf{x}), y\} d\mathbf{x}.$$

Then, q is increasing with $q(0) = 0$. If f is unbounded, then q is strictly increasing for all y with $\lim_{y \rightarrow \infty} q(y) = 1$. If f is bounded such that $\max(f) = F$, then q is strictly increasing for $0 \leq y \leq F$ and $q(y) = 1$ for all $y \geq F$.

We can rewrite f as a sum of two positive functions in the form

$$f(\mathbf{x}) = f^{(1)}(\mathbf{x}) + f^{(2)}(\mathbf{x}),$$

where $f^{(1)}(\mathbf{x}) = \min\{f(\mathbf{x}), Y\}$ and Y is positive. For a given $\epsilon > 0$, it is always possible to choose Y such that

$$1 - \frac{\epsilon}{4} < \int f^{(1)}(\mathbf{x}) d\mathbf{x} = q(Y) < 1,$$

and therefore

$$0 < \int f^{(2)}(\mathbf{x}) d\mathbf{x} < \frac{\epsilon}{4},$$

because q is continuous ranging between 0 and 1. The above “slicing” ensures that the function $f^{(1)}$ is bounded from above by Y . Let $h = Y/n$. Define a set of real numbers $\{r_{n,1}, \dots, r_{n,n}\}$ by

$$r_{n,j} = \sup\{r | p(r^2) \geq jh\}.$$

Here, p the non-increasing function defined on $\mathcal{R}^+ \cup \{0\}$ which satisfies $f(\mathbf{x}) = p[(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)]$. Setting

$$\omega_{n,j} = \frac{r_{n,j}^d}{\sum_{i=1}^n r_{n,i}^d},$$

we can then construct a density g_n such that

$$g_n(\mathbf{x}) = \sum_{j=1}^n \omega_{n,j} u_{n,j}(\mathbf{x}).$$

Next rewrite g_n as a sum of two positive functions in the form of

$$g_n(\mathbf{x}) = g_n^{(1)}(\mathbf{x}) + g_n^{(2)}(\mathbf{x}),$$

where

$$g_n^{(1)}(\mathbf{x}) = \sum_{j=1}^n r_{n,j}^d \cdot h \cdot u_{n,j}(\mathbf{x}).$$

Here, all three functions g_n , $g_n^{(1)}$ and $g_n^{(2)}$ are proportional to one another. Furthermore, by construction, $g_n^{(1)}$ is dominated everywhere by $f^{(1)}$. We also have

$$0 \leq f^{(1)}(\mathbf{x}) - g_n^{(1)}(\mathbf{x}) \leq \min\{f(\mathbf{x}), h\} \leq h.$$

It is therefore possible to choose n (and hence h) such that

$$\int |g_n^{(1)}(\mathbf{x}) - f^{(1)}(\mathbf{x})| d\mathbf{x} = \int \{f^{(1)}(\mathbf{x}) - g_n^{(1)}(\mathbf{x})\} d\mathbf{x} = q(h) < \frac{\epsilon}{4}.$$

Finally, applying the triangle inequality on integrals, we have

$$\begin{aligned} & \int |g_n(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} \\ & \leq \int |g_n^{(1)}(\mathbf{x}) - f^{(1)}(\mathbf{x})| d\mathbf{x} + \int g_n^{(2)}(\mathbf{x}) d\mathbf{x} + \int f^{(2)}(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

The first and third terms on the right-hand-side of the inequality are both less than $\epsilon/4$. The second term is

$$\int g_n^{(2)}(\mathbf{x}) d\mathbf{x} = 1 - \int g_n^{(1)}(\mathbf{x}) d\mathbf{x} < 1 - \int f_n^{(1)}(\mathbf{x}) d\mathbf{x} + \frac{\epsilon}{4} = \frac{\epsilon}{2}.$$

Hence, we arrive at

$$\int |g_n(\mathbf{x}) - f(\mathbf{x})| d\mathbf{x} < \epsilon,$$

completing the proof of Theorem 2.2. \square

ACKNOWLEDGEMENTS

This paper results from the Ph.D. work of Nicholas Chia at the Institute of Statistical Mathematics. Nicholas Chia would like to express his gratitude to John Copas, Shinto Eguchi, Katuomi Hirano, Satoshi Kuriki, Kunio Shimizu and Yoshiyasu Tamura for their kind and helpful advices. Bill Farebrother graciously supplied the proof of Theorem 5.1, which was crucial to the proof of Lemma 2.1 and subsequently Theorem 2.3. Nicholas Chia dedicates this paper to the memory of Taichi Morichika, who is outlived only by his creativity and passion in research.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, (6): 716–723.
- Anderson, T.W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities, *Proc. Amer. Math. Soc.*, **6**, 170–176.
- Asuncion, A. and Newman, D.J. (2007). UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- Berkhin, P. (2002). Survey of clustering data mining techniques, *Digital paper available from the internet*.
- Bernstein, D.S. (2005). *Matrix Mathematics*, Princeton University Press, Princeton, Oxford.
- Chia, N. and Nakano, J. (2009). \mathcal{M} -decomposability and symmetric unimodal densities in one dimension, *Ann. Inst. Stat. Math.* **61**(2), June 2009, 275–289.
- Comon, P. (1994). “Independent Component Analysis, a new concept?”, *Signal Processing, Elsevier, Special issue on Higher-Order Statistics*, **36**(3), April 1994, 287–314.
- Cover, T.M. and Thomas, J.A. (1988). Determinant inequalities via information theory, *SIAM J. Matrix Anal. Appl.*, **9**, No. 3, July 1988, 384–392.
- Dharmadhikari, S.W. and Joag-Dev, K. (1987). *Unimodality, Convexity and Applications*, Academic Press, New York.
- Duda, R.O., Hart, P.E. and Stork, D. G. (2001). *Pattern Recognition*, Second Edition Wiley-Interscience, New York.
- Fang, K.T., Kotz, S. and Ng, K.W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London.
- Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components, *Statistics and Computing*, **14**, 11–21.
- Farebrother, R.W. and Wrobel, I. (2002). Regular and reflected rotation matrices, *IMAGE*, **29**, 2002, 24–25.
- Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*, The MIT Press.
- Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*, Springer-Verlag, Berlin.
- Hardy, G., Littlewood, J.E. and Pölya, G. (1988). *Inequalities*, (Second Edition), Cambridge University Press, Cambridge.

- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning* Springer-Verlag, New York.
- Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications, *Neural Networks*, **13**(4-5): (2000), 411–430.
- Ibragimov, I.A. (1956, in Russian). On the composition of unimodal distributions, *Theor. Probability Appl.*, **1** (1956), 255–260.
- Kotecha, J.H. and Djurić, P.M. (2003). Gaussian particle filtering, *IEEE Transactions on Signal Processing*, **51**, No. 10, 2592–2601.
- Kotz, S., Read, C.B., Balakrishnan, N., Vidaković, B. (2005). *Encyclopedia of Statistical Sciences*, (16 Volume Set, Second Edition), John Wiley and Sons.
- Lee, D.S. and Chia, N. (2002). A particle algorithm for sequential Bayesian parameter estimation and model selection, *IEEE Transaction on Signal Processing*, **50**, No. 2, Feb 2002, 326–336.
- McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.
- McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, Wiley Interscience, New York.
- Pölya, G. and Szegő, G. (1972). *Problems and Theorems in Analysis I*, (English Edition), Springer-Verlag, Berlin.
- Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components, *Journal of the Royal Statistical Society B*, **59**, No. 4, 731–792.
- Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, (Second Edition), Springer.
- Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.
- Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points (with discussion), *Journal of the American Statistical Association*, **85**, September 1990, No. 411, 633–651.
- Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley, New York.
- Shioda, R. and Tunçel, L. (2005). Clustering via minimum volume ellipsoids, *Research Report CORR 2005–12, Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada*.
- Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, Chapman and Hall, London.

THE INSTITUTE OF STATISTICAL MATHEMATICS, 10-3 MIDORI-CHO, TACHIKAWA, TOKYO 190-8562, JAPAN.

E-mail address: sha@ism.ac.jp (Nicholas Chia)

THE INSTITUTE OF STATISTICAL MATHEMATICS, 10-3 MIDORI-CHO, TACHIKAWA, TOKYO 190-8562, JAPAN.